Describing Variability of MSW Composition Data with the Log-logistic Distribution

Mark W. Milke*, Vincent Wong[1], Department of Civil Engineering, University of Canterbury, Private Bag 4800, Christchurch, 8140, New Zealand.

Edward A. McBean, School of Engineering, University of Guelph, Guelph, Ontario, Canada.

Abstract

Variations in solid waste composition data are necessary as inputs to solid waste planning, yet uncertainty exists regarding which probability distributions might be generally valuable to describe the variability. Twenty-two detailed analyses of solid waste from British Columbia, Canada, are fit to distributions using the BestFit software. Alternative distributions are ranked based on three goodness-of-fit parameters and twelve waste fractions. The log-logistic distribution is found to be the most able to fit over the wide range of composition types considered. The results are demonstrated to be insensitive to the number of waste components or to the choice of a two- or three-parameter distribution. Although other distributions are able to match better the waste composition for individual waste types, the log-logistic distribution is demonstrated to fit, overall, a wide variety of waste composition types.

Keywords: log-logistic distribution; solid wastes, composition data analysis; goodness-of-fit; BestFit.

---

* Contact Author: mark.milke@canterbury.ac.nz, ph 64-03-3642-248, fax 64-03-3642-758

[1] Currently at Christchurch City Council.

1. Introduction

Short-term (daily or truck-to-truck) variations in the composition of municipal solid waste can be important in a variety of planning and management situations. The variation can influence the choice of design capacity (for example, to store a recycled fraction at a materials recovery facility), or the risk of regulatory violation (for example, from a hazardous material introduced to an incinerator), or can direct advertising (to improve recycling efforts for specific areas). An understanding of the amount and type of variation is also needed to analyse a data set (for example, to set confidence bounds or to identify erroneous data points).

Two types of variations in solid waste composition are commonly seen in a histogram of data. First, many composition data sets are symmetrical about the mean, and the normal distribution is commonly used to describe data. However, this distribution requires a standard deviation smaller than the mean; in other words, the coefficient of variation (C.V.), which is the ratio of the standard deviation to the mean, has to be less than unity. The other type of variation is seen with data where the histogram has a long right-hand tail (also called positive skewness) and a relatively low mean. This distribution type can have a C.V. near or greater than unity. Many distributions can be used to describe data of this type, including the lognormal, extreme value, logistic, log-logistic, inverse Gauss, and Pearson distributions. Many of these distributions have multiple variations, the most common feature being a parameter that shifts the distribution without changing its shape.

Analysts can choose to pick a different probability distribution each time they need to describe variability in composition; however, this has disadvantages when attempting comparisons of variability between different solid waste components or over time. The use of multiple distribution types requires transformations between parameters, and the differing shapes of the distributions make it difficult to identify trends in variations even after transformations are made. Use of a single transformation for describing most of the common variations would make it easier for analysts to understand and explain the variability of data and their implications.

This research points to the potential broad use of the log-logistic function for describing variability in solid waste composition data. The log-logistic distribution has found use in a variety of disciplines. Ahmad et al. (1988), among others, fit river flows to the distribution for assessing floods. Gleaton and Lynch (2004) describe the strength of fibres with the distribution, while Gokhale and Khare (2005) use the distribution to describe variability in carbon monoxide concentrations at urban intersections. Calder et al. (2005) use the distribution to describe the distribution of times between rockfalls, and Kooijman (1987) uses it to describe the variability in toxicity between aquatic species.

The log-logistic distribution (Johnson, et al., 1995) can be written with a shape parameter, α, a scale parameter, β, and a shift parameter, γ. The probability density function for the distribution is:

$$f(x) = \alpha * [(x - \gamma)/\beta]^{\alpha-1} / \{\beta * \{1 + [(x - \gamma)/\beta]^{\alpha}\}^2\} \qquad (1)$$

The cumulative distribution function, the integral of the above over *x*, is:

$$F(x) = 1 / \{1 + [\beta / (x - \gamma)]^{\alpha}\} \tag{2}$$

The mean (for $\alpha > 1$), m_x, and standard deviation (for $\alpha > 2$), σ_x, of the variable are given by:

$$m_x = [\pi*\beta*\csc(\pi/\alpha) / \alpha] + \gamma \tag{3}$$

$$\sigma_x = \beta * \mathrm{sqrt}\{\pi * [2 \csc(2\pi/\alpha) - \pi \csc^2(\pi/\alpha)/\alpha] /\alpha\} \tag{4}$$

The log-logistic function has the facility to match both types of data seen in solid waste composition. Figure 1 shows equation (1) for a high positive skew variable with a large C.V. of 1, and for a symmetrical variable with a low C.V. of 0.1.

This paper compares the log-logistic distribution with other distributions that could be used to fit a wide variety of solid waste composition data.

2. Vancouver Data Set

The composition data set used for this research is for waste delivered at the Burnaby incinerator outside of Vancouver, British Columbia, Canada, in July and October of 1998. In 1997 the incinerator received 257,460 tonnes of municipal solid waste. These wastes were residential wastes and commercially collected wastes, of which much came to the incinerator from the North Shore Transfer Station. The communities from which the waste originated were suburban and moderately affluent, with blue-box kerbside collection of newsprint, metal and glass for recycling. In addition, the communities had recycling depots and beverage-bottle deposit laws.

Twenty trucks were sampled in July in proportion to the annual waste brought to the incinerator by each type of truck. Two additional trucks were sampled because of the availability of labour, giving a dataset with 22 values. For each selected truck, a portion of the contents was dumped on an asphalt surface. Then, representative parts of the portion were selected as per an ASTM method (ASTM, 1992). This was accomplished at successive representative, but random, locations throughout the dumped load. The resulting portions of waste were then placed into drums until a net sample of 136 kg was reached. The waste sample was first separated into eight major waste categories and placed in labelled containers and weighed. Waste in some of the eight major waste containers was further classified (e.g., paper waste was separated into office paper, newsprint, OCC, and other paper). The resulting dataset has data for 27 primary waste categories, many of which have numerous zero values and so were not analysed for this research. Some primary waste categories were further separated, giving data on a total of 36 categories. No statistically significant differences were found in the composition of wastes from the transfer station and direct haul vehicles. There were no trends over time in the data set, indicating that it is reasonable to assume independent samples.

An identical procedure was used in October of 1998 in a second study of waste composition. The second study focused on a few primary waste classifications, and

20 trucks were sampled. Tables 1 and 2 summarise the data used in this paper from the two sampling periods. Classifications with zero values have been excluded. The dataset contains variables with a range of means and coefficients of variation, which is of benefit when examining the general suitability of probability distributions to solid waste composition data.

3. Goodness-of-fit Analyses

This research uses the software BestFit 4.5 (Palisade Corporation, 2006) to fit distributions to data. The software uses a maximum likelihood estimation procedure to fit the parameters to the distributions requested (Devore, 2004). Fifteen continuous distributions were evaluated for goodness-of-fit (beta, exponential, extreme value, gamma, inverse gauss, logistic, log-logistic, lognormal, normal, Pareto, Pearson5, Pearson6, triangular, uniform, Weibull). The result of the fitting procedure is a set of fitted parameters and various statistics assessing the goodness-of-fit.

There is no single, universally best way to decide goodness-of-fit of data to a distribution. BestFit 4.5 provides three statistics for goodness-of-fit, all three of which are used here. The three are: the Chi-square statistic, the Kolmogorov-Smirnov (K-S) statistic, and the Anderson-Darling (A-D) statistic. For the Chi-square statistic, the range of data is divided into a number of bins, and the number of datapoints found in each bin is compared with the number of datapoints that is expected in that bin based on the fitted parameters. The resulting Chi-square statistic is given by:

$$\chi^2 = (N_1 - E_1)^2/E_1 + (N_2 - E_2)^2/E_2 + \ldots (N_k - E_k)^2/E_k \tag{5}$$

where N_i is the number of datapoints in bin i, E_i is the expected number of datapoints in bin i, and k is the number of bins. The value of the statistic depends slightly upon the choice of the number of bins and their size, which is a weakness of the use of this statistic for testing the goodness-of-fit for a continuous probability distribution. For this work, the default bins selected by BestFit are used.

The K-S statistic is the largest difference between the cumulative distribution of the data and of the fitted distribution. Because all cumulative distributions vary from 0 to 1, the statistic will tend to be small at the two extremes of the distribution. The implication is that the K-S statistic is better at discriminating poor fit near the mean, but worse at the extremes.

The A-D statistic is given by:

$$A = n^* \int \{[F(x) - F'(x)]^2 / [F'(x)*(1-F'(x))]\} f'(x)\, dx \tag{6}$$

where F(x) is the cumulative distribution of the actual data, f'(x) and F'(x) are the probability density and cumulative probability distributions for the fitted distribution, and n is the number of data points. In the A-D statistic, the difference between the two distributions is multiplied by a weighting factor that is larger at the two tails of the distribution when F(x) and F'(x) both approach either 0 or 1. In this way, the A-D statistic is better suited to discerning which distributions fit better at the extremes. All three are described in standard statistic textbooks (e.g., Johnson et al., 2004).

3.1 Analyses of Twelve Waste Fractions

The use of BestFit on the twelve waste components in Tables 1 and 2 resulted in a number of distributions that could fit individual components. Because of the many ways of defining goodness-of-fit, and because of varying goodness-of-fit determinations between components, it will never be possible to say that one distribution always fits the data better than others. For this study, we use a scoring system to evaluate the goodness-of-fit, giving three points for the distribution providing the best fit, two points for the distribution giving the second-best fit, one point for the distribution giving the third-best distribution, and no points for distributions giving poorer fits. With this system, a total of 216 points are assigned over the twelve components and three goodness-of-fit statistics. The results in Table 3 show the six highest-scoring distributions. The other nine distributions (e.g, normal) were also tried but did not score as highly. In a few cases one of the other nine distributions ranked in the top three on one of the three measures. The log-logistic distribution scored the best for these two datasets.

Figures 2 and 3 present sample results to show how the log-logistic distribution is able to fit a wide variety of data types relatively well. Figure 2 shows the paper composition data, which have a relatively high mean and low coefficient of variation. Both the normal and log-logistic distributions fit the data relatively well. Figure 3 shows the glass composition data, which have a relatively low mean and high coefficient of variation. Both the lognormal and log-logistic distributions fit the data relatively well. However, the normal distribution does not fit the glass distribution data well, and the lognormal distribution does not fit the paper distribution data well.

3.2 Analysis of sensitivity to coarseness of waste classification

The number of waste components employed in the Burnaby incinerator dataset is greater than the number found in many waste analysis studies. As a result, the means are likely to be lower than those found from other datasets and hence of interest was the question of the sensitivity of the results to the number of waste components employed.

To examine this, data for specific waste components were combined into four new hypothetical waste classifications. Four specific combinations were examined: paper+organics, paper+plastics, metals+glass+inorganics, and metals+glass. Table 4 shows the best-fit distributions for these four combinations of the waste components. The log-logistic distribution fits all combinations best, except for the combination of paper+organics.

For the Burnaby dataset, the paper+organics combination has a mean of 70% of the total waste stream, which is higher than all other means considered in this study. The reason that the log-logistic distribution cannot fit these data well is because it is not able to represent negative skew (left-hand tail). The data for this combination have a slight negative skew (skewness coefficient of -0.06). The result for paper+organics highlights a limitation in the use of the log-logistic function: it can be expected to fit solid waste composition distributions worse as the data exhibit more negative skew (which is more likely as the mean increases above 50%).

3.3 Analysis of sensitivity to number of fitted parameters in distribution

The distributions listed in Table 3 contain some distributions with two fitted parameters, and others, including the log-logistic, with three. One criterion for choosing a distribution is the desire to fit fewer parameters. An additional analysis was conducted examining the sensitivity of goodness-of-fit to the number of parameters. The two-parameter log-logistic function (with the shift parameter, γ in equation (1), set to 0) was fitted to the twelve waste components, and the median K-S and A-D statistics found are given in Table 5. Because the Pearson 5 distribution with three parameters also showed relatively good fit in Table 3, the two-parameter version of this distribution was also selected for analysis. The results are shown alongside that of the three-parameter log-logistic distribution.

The reduction from three parameters to two parameters would be expected to result in worse fits, and worse fits can be seen by greater K-S and A-D statistics. Table 5 shows slightly worse fit for the two-parameter log-logistic distribution compared to the three-parameter form. The two-parameter Pearson 5 distribution appears to fit the data worse than the two-parameter log-logistic distribution, just as was found with their three-parameter forms in Table 3. The results indicate that the two-parameter form of the log-logistic distribution could be a suitable generalised distribution for solid waste composition data when a desire exists to reduce the number of parameters.

4. Discussion

The common techniques used to estimate the number of samples needed to reach a specified precision in solid waste composition assume a normal distribution for the data (Sharma and McBean, 2006; Sfeir et al., 1999; Leroy et al., 1992). Estimates of the means are commonly insensitive to the choice of the underlying distribution, although this assumption could be important when estimating the standard deviation or unlikely events. For example, when an unusual data point is found, it is common to use techniques to estimate the likelihood of an outlier. Outlier identification procedures often require the assumption of a probability distribution and so the conclusion of whether the data point is erroneous or not can depend upon the assumed distribution (Gilbert, 1987; McBean and Rovers, 1998). Future research into techniques for analysis of solid waste composition data using the log-logistic distribution would appear needed..

Although this analysis shows the potential for the log-logistic distribution to fit a wide variety of solid waste composition data, it is important to appreciate the limitations of this analysis. This study considers datasets with non-zero values. There are many solid waste composition datasets where measured components have zero composition in certain samples (e.g. car batteries). In these cases, it might be more appropriate to use a mixed probability distribution. In addition, this study examines only one dataset collected in a particular fashion, and it is possible that results will not be readily transferable.

The scoring system used here is relatively simplistic and the sensitivity of the results to the choice of a scoring system has not been analysed. The degree to which a given rank is better than another is ignored. Although similar ranking approaches are commonly used in non-parametric statistics (McBean and Rovers, 1998), other approaches could also be used. The data for all waste components are assumed to be of equal value to a decision-maker, and the approach described herein gives them each equal weight. It also ignores the distinctions seen between the measures of goodness-of-fit, and so ignores the potential to give more weight to one or another.

Further research with additional data sets could indicate that particular distributions are suited to particular waste types or data applications. This research has focused on the potential for finding one distribution to fit a wide variety of waste composition datasets.

It is easier to accept the assumption of a distribution when there is a plausible mechanism that might explain why the data would tend towards a particular distribution. The authors cannot provide any theoretical, underlying reason why solid waste composition data should fit a log-logistic distribution. One particular weakness of the log-logistic distribution is that it is not bounded at 100%, and hence there will always be a very small probability that over 100% of a component is found in the sample, which is inappropriate. Although the Beta distribution was not able to fit the data analysed here better than the log-logistic distribution, there might be other bounded distributions that could provide better fits. In any case, it is important that those who might use the log-logistic distribution check for the likelihood that the fitted distribution could indicate non-zero probabilities for percentages greater than 100%.

The use of one probability distribution to analyse trends in multiple solid waste composition data sets could have value to waste managers. These trends could exist in time or in space. It is common to track specific solid waste parameters using periodic sampling. The analysis of the data in one instance with one distribution, and in another instance with a second will make it difficult to compare between data sets. Although it could be more difficult for a waste manager to use and to explain a log-logistic distribution than a normal or lognormal one, this research indicates that one less well known distribution could provide a good fit in a wide variety of situations. There are potential benefits of using a less well known distribution, and this should be considered, along with ease of use, when analysing data. Further research is needed into the use of one distribution to analyse trends in solid waste composition data.

5. Conclusion

This research shows that although other distributions are able to match better individual data, the flexibility of the log-logistic distribution makes it able to fit a wide variety of solid waste composition data relatively well and, overall, better than other distributions.

Acknowledgements

Support from the Royal Society of Edinburgh, the University of Canterbury Erskine Fund, and the Canada Research Chairs program are gratefully acknowledged.

References

Ahmad, M.I., Sinclair, C.D., & Werritty, A. (1998) Log-logistic flood frequency analysis. *Journal of Hydrology*, **98**, 205-224.

ASTM (1992) "Standard Test Method for the Determination of the Composition of Unprocessed Municipal Solid Waste", D523 I-92.

Calder, E.S., et al. (2005) Probabilistic analysis of rockfall frequencies during an andesite lava dome eruption: The Soufriere Hills Volcano, Montserrat. *Geophysical Research Letters*, **32**, n 16, L16309.

Devore, J. (2004) *Probability and Statistics for Engineering and the Sciences, Sixth edition.* Thomson.

Gilbert, R.O. (1987) *Statistical Methods for Environmental Pollution Monitoring*. Van Nostrand Reinhold.

Gleaton, J.U., & Lynch, J.D. (2004) On the distribution of the breaking strain of a bundle of brittle elastic fibers. *Advances in Applied Probability*, **36**, 98-115.

Gokhale, S., & Khare, M. (2005) A hybrid model for predicting carbon monoxide from vehicular exhausts in urban environments. *Atmospheric Environment*, **39**, 4025-40.

Johnson, N.L, Kotz, S., & Balakrishnan, N. (1995) *Continuous Univariate Distibutions, Volume 2*, 2nd edn. J Wiley.

Johnson, R.A., Miller, I., & Freund, J. (2004) *Miller and Freund's Probability and Statistics for Engineers*, 7th edn. Prentice Hall.

Kooijman, S.A.L.M. (1987) Safety Factor for LC//5//0 values allowing for differences in sensitivity among species. *Water Research*, **21**, 269-276.

Leroy, D., Giovannoni, J.-M., & Maystre, L.-Y. (1992) Sampling method to determine a household waste composition variance. *Waste Management and Research*, **10**, 3-12.

McBean, E.A., & Rovers, F.A. (1998) *Statistical Procedures for Analysis of Environmental Monitoring Data and Risk Assessment*. Prentice-Hall.

Palisade Corporation (2007) Best Fit 4.5, http://www.palisade.com.au/bestfit/ (January 23, 2008)

Sfeir, H., Reinhart, D.R., & McCauley-Bell, P.R. (1999) An evaluation of municipal solid waste composition bias sources. *Journal of the Air and Waste Management Association*, **49**, 1096-1102.

Sharma M., & E. McBean (2006) A methodology for solid waste characterization based on diminishing marginal returns. *Waste Management*, **27**, 337-344.

Table Captions

Table 1. Waste composition data from 22 samples at the Burnaby incinerator, July, 1998.

Table 2. Waste composition data from 20 samples at the Burnaby incinerator, October, 1998.

Table 3. Scored assessment of goodness-of-fit of six distributions (with number of parameters in brackets) to solid waste composition data from Burnaby, Canada. Each entry gives three scores for goodness-of-fit measured by Chi-square, Kolmogorov-Smirnov, and Anderson-Darling statistics respectively. A score of 3 represents the best fitted distribution, 2 the second-best, 1 third-best, and 0 is assigned for worse than the third-best distribution.

Table 4. Goodness-of-fit of five distributions to combinations of solid waste composition data from Burnaby, Canada. Each entry gives three scores for goodness-of-fit measured by Chi-square, Kolmogorov-Smirnov, and Anderson-Darling statistics respectively. A score of 3 represents the best fitted distribution, 2 the second-best, 1 third-best, and 0 is assigned for worse than the third-best distribution.

Table 5. Median Kolmogorov-Smirnov and Anderson-Darling statistics over the twelve waste components for selected two- and three-parameter distributions.

Figure Captions

Figure 1. The varied shapes of the log-logistic distribution. (a) a skewed form with mean and standard deviation of 2 (beta =1.6; alpha = 2.7); (b) a symmetric form with mean of 20 and standard deviation of 2 (beta = 20; alpha = 18).

Figure 2. Paper percentage in 22 solid waste samples taken at the Burnaby incinerator, BC, and fit of the data using BestFit to (a) the normal distribution, and (b) the shifted log-logistic distribution.

Figure 3. Glass percentage in 22 solid waste samples taken at the Burnaby incinerator, BC, and fit of the data using BestFit to the (a) unshifted lognormal distribution, and (b) shifted log-logistic distribution.

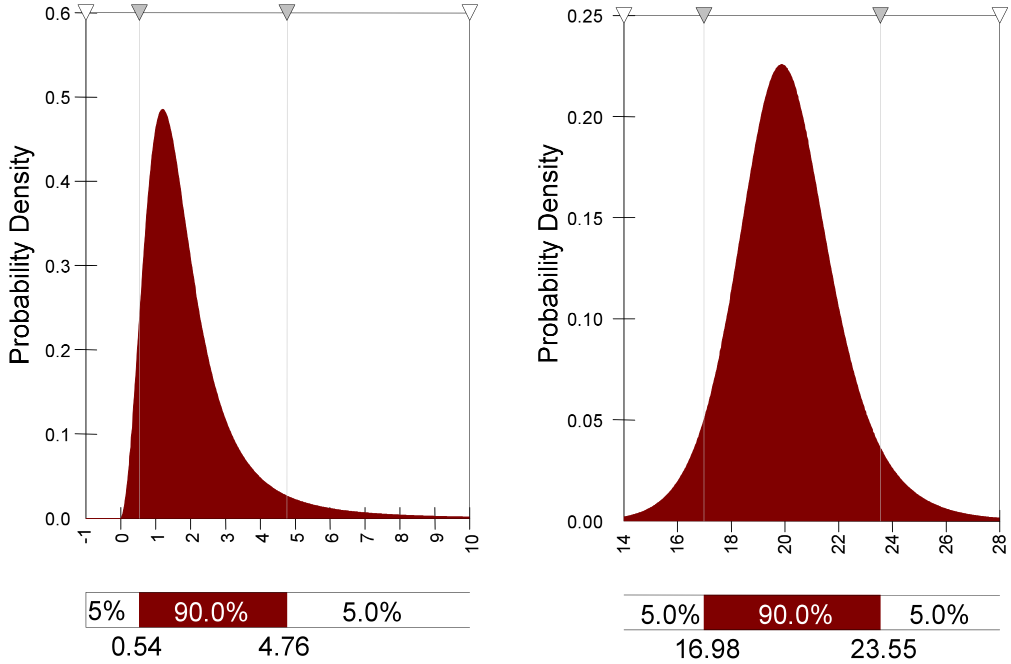


Fig. 1: The varied shapes of the log-logistic distribution. (a) a skewed form with mean and standard deviation of 2 (β =1.6; α = 2.7); (b) a symmetric form with mean of 20 and standard deviation of 2 (β = 20; α = 18).

Major Waste Component	Mean (%)	Standard Deviation (%)
Organics	37.4	11.1
Paper	32.3	10.6
Plastics	13.3	5.4
Household Hazardous	5.9	3.0
Metals	3.4	1.5
Glass	3.1	2.3
Inorganic	2.9	3.8
Fines	1.2	1.7

Table 1: Waste composition data from 22 samples at the Burnaby incinerator, July, 1998.

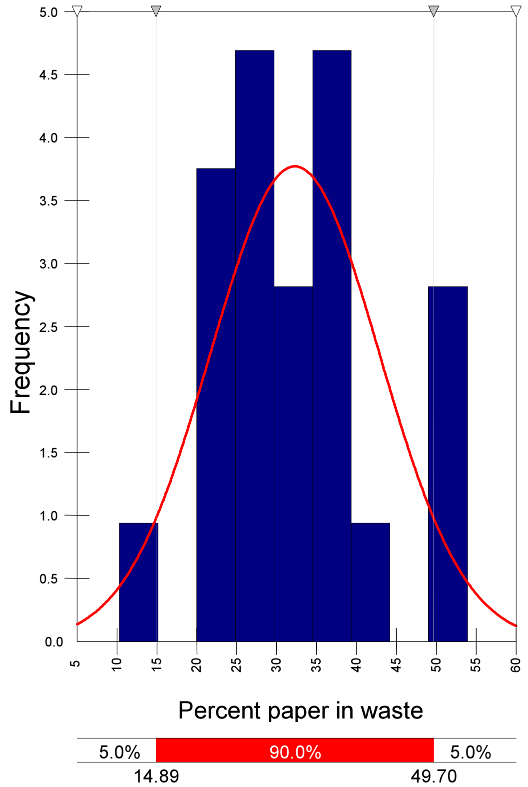
Primary Waste Component	Mean (%)	Standard Deviation (%)
OCC Paper	5.22	3.48
Film Plastic	6.70	1.72
Plastic (not Film or PVC)	3.43	2.21
Yard and Garden (not grass)	4.55	5.33

Table 2: Waste composition data from 20 samples at the Burnaby incinerator, October, 1998.

Waste Component	Sample Set	Extreme Value [2]	Inverse Gauss [3]	Logistic [2]	Log-logistic [3]	Lognormal [3]	Pearson 5 [3]
Organics	July, 98	2/0/0	0/1/1	0/0/0	3/3/3	1/0/0	0/2/2
Paper	July, 98	2/0/0	0/0/1	0/0/0	3/3/0	0/1/2	0/2/3
Plastics	July, 98	3/0/2	2/2/0	1/0/0	0/1/1	0/0/0	0/3/3
Household Hazardous	July, 98	0/0/0	0/0/0	2/3/3	0/2/2	0/1/0	0/0/0

Metals	July, 98	3/2/2	0/0/0	0/0/0	1/3/3	0/0/0	0/1/0
Glass	July, 98	3/0/2	2/0/1	1/0/3	0/1/0	0/2/0	0/3/0
Inorganics	July, 98	3/1/1	0/3/2	2/0/0	0/0/0	0/0/0	0/0/0
Fines	July, 98	0/0/0	1/2/2	0/0/0	0/0/0	0/0/0	2/3/3
OCC Paper	Oct., 98	3/0/0	2/0/0	0/0/0	0/0/2	1/1/1	0/2/3
Film Plastic	Oct., 98	1/0/0	0/0/0	0/2/2	0/3/3	0/0/0	0/1/1
Plastic (not film or PVC)	Oct., 98	0/0/0	2/3/1	0/0/0	1/1/3	0/2/2	0/0/0
Yard and Garden (not grass)	Oct., 98	2/0/2	0/1/0	0/0/0	0/0/0	1/3/1	0/0/0
Total Score		34	29	19	42	19	34

Table 3: Overall weighted assessment of Goodness-of-Fit of six distributions (with number of parameters in brackets) to solid waste composition data from Burnaby, Canada. Each entry gives three scores for goodness-of-fit measured by Chi-squared, Kolmogorov-Smirnov, and Anderson-Darling statistics respectively. A score of 3 represents the best fit distribution, 2 the second-best, 1 third-best, and 0 is assigned for worse than the third-best distribution.



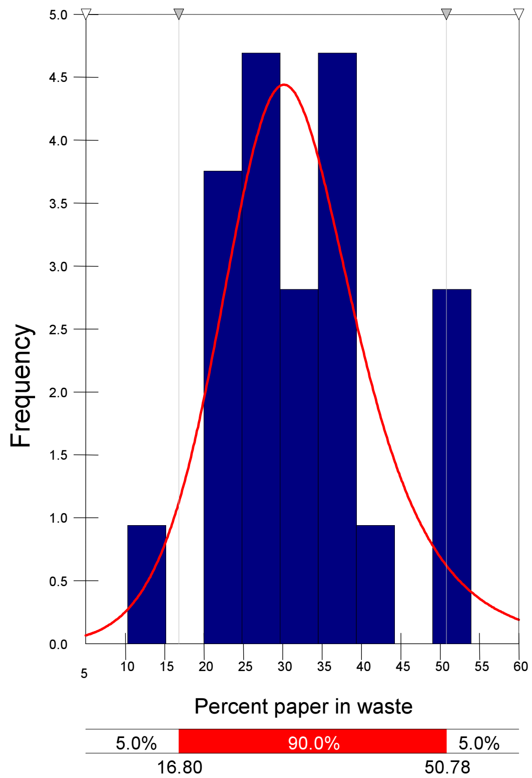
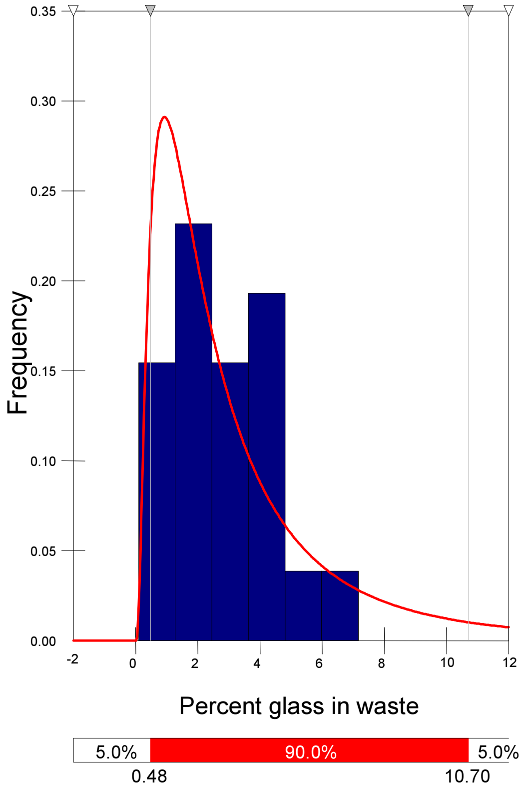


Fig. 2: Paper percentage in 22 solid waste samples taken at the Burnaby incinerator, BC, and fit of the data using BestFit to (a) the normal distribution, and (b) the shifted log-logistic distribution.



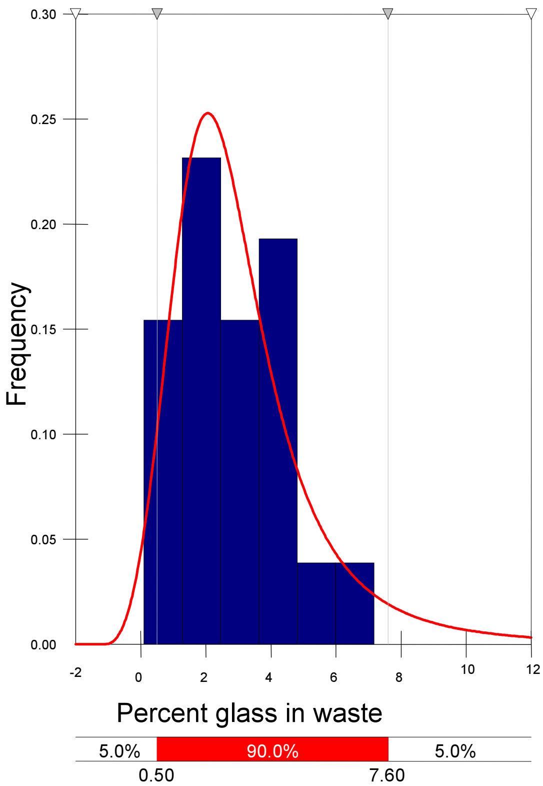


Fig. 3: Glass percentage in 22 solid waste samples taken at the Burnaby incinerator, BC, and fit of the data using BestFit to the (a) unshifted lognormal distribution, and (b) shifted log-logistic distribution.

Waste Combination	Mean	Inverse Gauss	Logistic	Log-logistic	Normal	Pearson 5
Paper + Organics	69.7	0/0/0	1/3/3	0/0/0	3/2/2	0/0/0
Paper + Plastics	45.6	3/0/1	2/2/2	1/3/3	0/0/0	0/0/0
Metals + Glass + Inorganics	9.4	1/0/0	3/0/0	0/3/3	0/0/0	0/2/2
Metals + Glass	6.5	3/0/0	2/2/2	1/3/3	0/0/0	0/1/1

Table 4: Goodness-of-fit of five distributions to combinations of solid waste composition data from Burnaby, Canada. Each entry gives three scores for goodness-of-fit measured by Chi-squared, Kolmogorov-Smirnov, and Anderson-Darling statistics respectively. A score of 3 represents the best fit distribution, 2 the second-best, 1 third-best, and 0 is assigned for worse than the third-best distribution.

Distribution	No. Parameters	Median Kolmogorov-Smirnov Statistic	Median Anderson-Darling Statistic
Log-logistic	2	0.125	0.412
Pearson 5	2	0.184	3.34
Log-logistic	3	0.118	0.292

Table 5: Median Kolmogorov-Smirnov and Anderson-Darling statistics over the twelve waste components for selected two- and three-parameter distributions.